INRA Prod. Anim.,
2010, 23 (4), 343-358

Apports de la génomique fonctionnelle à la cartographie fine de QTL

G. LE MIGNON[1,2,3], Y. BLUM[1,2,4], O. DEMEURE[1,2], C. DIOT[1,2], E. LE BIHAN-DUVAL[5], P. LE ROY[1,2], S. LAGARRIGUE[1,2]

[1] INRA, UMR598 Génétique Animale, F-35000 Rennes, France

[2] Agrocampus Ouest, UMR598 Génétique Animale, F-35000 Rennes, France

[3] ITAVI, 28 rue du Rocher, F-75008 Paris, France

[4] Agrocampus Ouest, Laboratoire de Mathématiques Appliquées, F-35000 Rennes, France

[5] INRA, UR83 Recherches Avicoles, F-37380 Nouzilly, France

Courriel : Sandrine.Lagarrigue@agrocampus-ouest.fr

De nombreux programmes de recherche en génétique animale ont permis de localiser des régions QTL alors que les mutations causales sous-jacentes sont encore rarement identifiées. Après avoir introduit le concept de QTL d'expression, cet article présente les principales stratégies utilisant des données d'expression génique pour mieux caractériser ces régions QTL et en faciliter ainsi l'identification des mutations causales.

De nombreux programmes de recherche en génétique animale ont pour objectif de localiser des QTL (*Quantitative Trait Locus),* régions du génome responsables de la variabilité de caractères complexes d'intérêt économique et d'en identifier le(s) polymorphisme(s) causal(aux) sous-jacent(s). Depuis les années 2000, des technologies de génomique fonctionnelle se sont développées permettant de mesurer simultanément l'expression de l'ensemble des gènes d'un génome. Ces phénotypes plus élémentaires peuvent être des quantités d'ARN messager (ARNm), de protéines ou par extension de métabolites. Nous proposons dans cet article de présenter les principales stratégies utilisant des données d'expression dans le cadre de la cartographie de QTL. L'une de ces stratégies que nous appellerons «décomposition du caractère» peut se révéler efficace pour affiner voire même permettre la détection de nouveaux QTL. Une seconde stratégie, plus courante et appelée «eQTL» (pour QTL d'expression), peut apporter de nouvelles informations fonctionnelles sur la région QTL. D'autres approches variantes de la précédente peuvent contribuer à une localisation plus fine de la région QTL.

Dans un premier temps, nous introduirons le concept de QTL d'expression (eQTL) et les principaux résultats de cartographie d'eQTL que l'on peut extraire de la bibliographie indépendamment du contexte de recherche de QTL (parties 1 et 2). Nous exposerons ensuite les principales stratégies utilisant des données transcriptomiques dans le cadre de la détection de QTL (partie 3).

1 / QTL d'expression

1.1 / Principes généraux

Un gène est une séquence du génome qui est exprimée et va, dans un bon nombre de cas, conduire à la synthèse d'une ou de protéines qui lui sont spécifiques, les protéines représentant des entités fonctionnelles de la cellule. On estime aujourd'hui que les génomes des animaux d'élevage sont composés d'environ 40 000 gènes, eux-mêmes à l'origine de plusieurs millions de protéines (http://www.ensembl.org/species/Info/StatsTable?db=core). En raison de leurs fonctions très variées les protéines jouent un rôle majeur dans l'établissement des caractères d'intérêt agronomiques visibles à l'échelle de l'animal. Certaines sont des protéines de structure, d'autres des enzymes catalysant des réactions biochimiques, d'autres encore sont des régulateurs des différents mécanismes cellulaires permettant le décodage de l'information génétique depuis les gènes situés dans le noyau de nos cellules jusqu'aux protéines qu'ils codent. En effet, le passage du gène à la protéine est un processus complexe comprenant deux étapes majeures (encadré 1). La première étape, la transcription, consiste à transcrire le gène en un certain nombre de copies, dites transcrits ou ARN messager. Selon le nombre de copies générées, un gène sera dit plus ou moins exprimé dans un tissu considéré. La seconde étape, la traduction, est le décodage de chaque copie d'ARNm en protéine. A chaque étape, transcrits et protéines peuvent être dégradés.

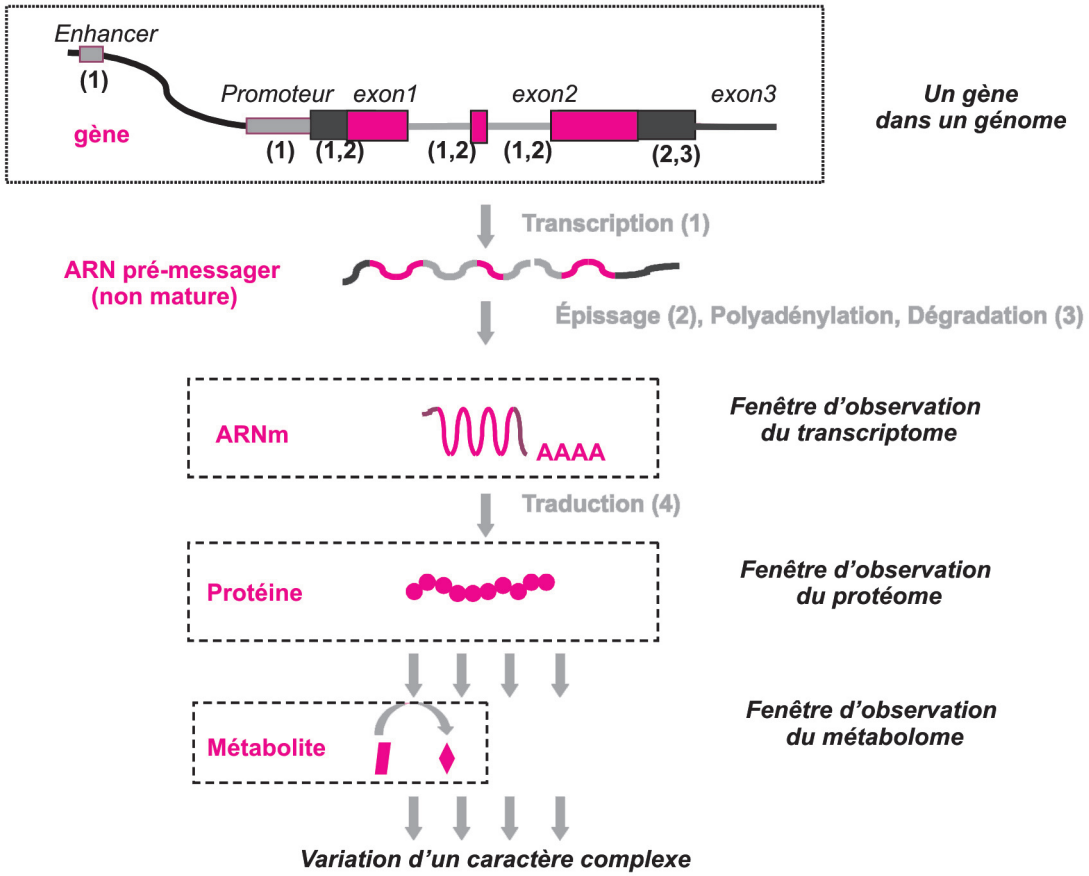
Ainsi, la mesure de la quantité d'un transcrit (ou ARNm) ou d'une protéine résulte des mécanismes de synthèse mais aussi de dégradation de ces molécules. A cause de leur complexité, les voies de contrôle par la cellule de ces mécanismes sont encore loin d'être élucidées. Aussi, la quantité d'ARNm d'un gène donné dans un tissu (ou des protéines associées) est probablement régulée par de nombreuses protéines et donc de gènes les codant.

Etant donné sa structure, un gène peut se décomposer en différentes régions. Les régions codantes (boîtes rouges sur l'encadré 1) portent l'information de la future protéine. Les autres régions dites régulatrices (boîtes noires ou grises ou traits gris sur l'encadré 1) participent au contrôle de la transcription du gène en ARNm (*enhancer*, promoteur, introns…) ou à la stabilité/dégradation de ces ARNm (régions transcrites mais non codantes en fin de gène).

Encadré 1. *Rappel des différentes étapes de transmission de l'information du gène à la protéine correspondante.*

Exons (régions transcrites dans l'ARNm) :
Régions codantes (traduites en protéine)
Régions non codantes (non traduites en protéines)
Introns (régions inter-exoniques, excisées de l'ARNm pré-messager)

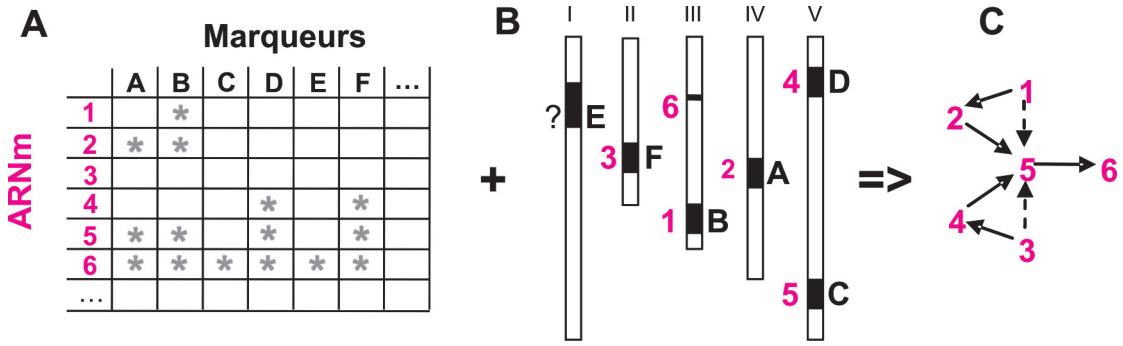
(1), (2) et (3) : Régions du gène intervenant dans la régulation de sa transcription (1), dans l'épissage de ses ARNm non matures (2) ou encore dans la stabilité de ses ARNm (3)



Les gènes sont dans un premier temps transcrits en ARN pré-messagers. La maturation des pré-messagers implique principalement l'excision des séquences introniques (on parle alors d'épissage) et l'addition d'une succession de plusieurs ribonucléotides de types Adénosine (queue poly A) en extrémité 3' de la molécule (ou polyadénylation). L'ARN acquiert alors une meilleure stabilité, nécessaire pour limiter les phénomènes de dégradation dus en majeure partie aux enzymes de types RNAse. C'est à ce stade, en aval des nombreux mécanismes ci-dessus mentionnés, que l'on peut mesurer le niveau d'expression d'un gène. Cet ARNm mature est ensuite traduit en protéine lors de la traduction. Certaines protéines, les enzymes, peuvent conduire à la transformation de métabolites. Un ensemble de protéines contribue par la suite aux variations d'un caractère complexe visible à l'échelle de l'animal.

Aujourd'hui, grâce aux technologies de la génomique fonctionnelle, il devient possible de mesurer simultanément dans un tissu donné l'expression de plusieurs gènes au niveau de leur ARNm (transcriptome) ou de leurs protéines (protéome). Il est également possible de mesurer différentes métabolites d'un tissu (métabolome).

Figure 1. *Interprétation conjointe des positions chromosomiques des régions eQTL et des gènes qu'elles contrôlent en vue de reconstituer une voie métabolique (d'après Jansen et Nap 2001).*



L'identification de régions eQTL (au niveau des marqueurs A, B, C,…) régulant des gènes (gène 1, 2, 3…) (indiqué en A) couplée à la localisation de ces gènes sur le génome, en particulier dans les régions eQTL(B) permet de reconstituer en théorie une voie métabolique ou voie de régulation (C). Par exemple, l'expression du gène 6 est contrôlée par les 6 régions indiquées, il est donc probable que ce gène code pour une protéine en fin de voie métabolique ou de régulation. Le gène 5 partageant 4 de ces 6 régions et se trouvant être localisé dans la région C contrôlant le gène 6, serait donc régulateur du gène 6. Par un raisonnement similaire généralisé à l'ensemble des gènes et régions indiquées à gauche de la figure, une voie métabolique ou de régulation est proposée à droite de la figure.

Une mutation dans les régions régulatrices d'un gène peut donc conduire à une variation de la quantité de ses transcrits, une mutation dans la région codante peut affecter la fonctionnalité de la protéine codée, et éventuellement affecter les quantités d'ARNm d'autres gènes. Dans ce cas, une mutation dans un gène donné peut avoir des impacts sur la quantité d'ARNm d'autres gènes sans que la quantité de ses propres ARNm ne soit affectée.

Grâce à l'essor des technologies de génomique fonctionnelle dans les années 2000, il est devenu possible de quantifier dans un tissu donné les niveaux d'ARNm de l'ensemble des gènes contenus dans un génome. Ces niveaux de transcrits constituent le «transcriptome» du tissu considéré. De même, il est possible de quantifier quelques centaines de protéines en une seule expérimentation (appelé alors protéome) ou quelques dizaines à centaines de métabolites (appelé alors métabolome) (encadré 1). Concernant les métabolites, une partie d'entre eux sont le produit de réactions biochimiques effectuées par des proteines (et donc plus en amont par des gènes codant ces protéines). Aussi, la fenêtre d'observation des métabolites est intéressante à la fois pour préciser un phénotype d'intérêt agronomique mais aussi pour apprécier l'activité fonctionnelle/expressionnelle d'un génome. L'accès à ces quantités d'ARNm, de protéines ou de métabolites permet donc d'observer à grande échelle les phénotypes intermédiaires entre les polymorphismes du génome et les caractères d'intérêt agronomique. Il est intéressant de noter que les technologies du transcriptome sont les seules à permettre l'analyse de l'expression de l'ensemble des gènes d'un génome. Aussi sont-elles plus utilisées que les autres technologies du protéome ou du métabolome. Les références bibliographiques et les illustrations mentionnées dans cet article seront donc centrées sur les seules données du transcriptome.

Selon les mêmes principes que ceux concernant la cartographie de QTL classique pour des caractères visibles au niveau de l'animal, il devient maintenant possible de détecter des régions du génome contrôlant le niveau d'ARNm d'un ou de plusieurs gènes. Ces régions sont appelées eQTL pour QTL d'expression. Le niveau d'ARNm d'un gène est alors considéré comme un caractère complexe à part entière. En 2001, Jansen et Nap proposent de nommer ce nouveau concept «*genetical genomics*» (pour génétique de la génomique ou encore «génétique génomique») pour lequel une analyse de la liaison génétique entre marqueurs et expression d'un gène permet de mettre en évidence des locus responsables d'une part de la variation de son expression. Comme nous le verrons plus loin, il est ainsi possible d'identifier des gènes dont les expressions sont affectées par plusieurs locus ou encore un ensemble de gènes partageant les mêmes locus de contrôle. Comme indiqué dans la figure 1, l'analyse de ces différentes liaisons génétiques entre locus et gènes régulés combinées à la localisation des gènes dans le génome devrait en théorie permettre selon Jansen et Nap de reconstituer des voies de régulation ou voies métaboliques en mettant en évidence le gène le plus aval de la voie, celui dont l'expression est gouvernée par le plus grand nombre de locus.

La «génétique génomique» constitue une nouvelle manière d'observer les événements de régulation génique à une échelle encore jamais explorée.

Un an plus tard, le concept de la «génétique génomique» est validé par des premières expériences chez la levure où une cartographie de régions contrôlant la variabilité de l'expression de gènes est réalisée à l'échelle du génome (Brem *et al* 2002). Depuis, de nombreuses études ont été menées visant à cartographier les eQTL de gènes chez la levure (Yvert *et al* 2003), chez des espèces modèles comme la souris (Schadt *et al* 2003), la drosophile (Wayne et McIntyre 2002), le rat (Hubner *et al* 2005) ou encore chez l'Homme (Monks *et al* 2004, Morley *et al* 2004). On note également des études similaires chez les végétaux comme le maïs (Schadt *et al* 2003), l'*Eucalyptus* (Kirst *et al* 2004), *Arabidopsis thaliana* (DeCook *et al* 2006) et l'orge (Potokina *et al* 2008).

Le concept de la cartographie de régions eQTL peut également être utilisé pour des caractères de type «quantité d'une protéine (pQTL)» (Zivy et de Vienne 2000) ou encore «quantité d'un métabolite (mQTL)» (Ferrara *et al* 2008). Les premiers travaux visant à cartographier des régions responsables de la variabilité d'un taux protéique sont d'ailleurs plus anciens que la cartographie de QTL d'expression au niveau des ARNm (De Vienne *et al* 1999).

Comparé à un phénotype d'intérêt agronomique, les phénotypes élémentaires que sont les transcrits ou les protéines d'un gène diffèrent par le nombre plus faible de mécanismes impliqués dans leur variation. Malgré tout, ces mécanismes sont divers comme indiqué plus haut (mécanismes de contrôle de l'activité transcriptionnelle du gène ou de la dégradation de ces ARNm ou protéines). Ainsi, la variation des quantités d'ARNm d'un gène peut être la résultante d'un polymorphisme présent

dans le gène lui-même ou présent dans un autre gène qui serait alors impliqué dans l'un des mécanismes de contrôle précédemment cité. Le vocabulaire qualifiant les régions eQTL s'est donc enrichi par rapport à celui qualifiant les QTL.

1.2 / Les *cis* et *trans* eQTL

Une région eQTL sera qualifiée de «*cis* eQTL» si sa localisation est proche de la position du gène dont elle gouverne l'expression. Au contraire, si ce gène est positionné ailleurs dans le génome par rapport à la position de la région eQTL, celle-ci sera qualifiée de «*trans* eQTL». Ce vocabulaire de *cis* et *trans* eQTL a été emprunté au domaine de la biologie moléculaire. En biologie moléculaire, le terme *trans* est souvent associé aux facteurs de transcription régulant un gène et se fixant en *trans* sur des séquences en général promotrices du gène régulé appelées, elles, séquences-*cis*. Ces termes *trans* et *cis* ont donc une connotation mécanistique. Ils ne sont donc pas toujours appropriés aux eQTL, les études de cartographie ne renseignant pas les mécanismes moléculaires mis en jeu dans les régulations. Certains auteurs estiment préférable d'adopter les termes d'eQTL «locaux» et «distants» (figure 2) faisant référence à la position des marqueurs génétiques par rapport à la position des gène régulés (Rockman et Kruglyak 2006) : le terme d'eQTL local est communément employé lorsque la position du marqueur est comprise dans une fenêtre de taille arbitraire (le plus souvent 5 à 10 Mb) autour de la position du gène régulé, dans le cas contraire, on parle d'eQTL «distant».

D'après les connaissances que nous avons de la régulation des niveaux d'ARNm d'un gène, les *cis* eQTL peuvent être causés par des polymorphismes dans les régions régulatrices des gènes eux mêmes : dans leurs séquences promotrices, introniques, exoniques non codantes, ou *enhancer* (cf. encadré 1). Cependant, la démonstration de l'effet d'un polymorphisme dans des régions régulatrices sur l'expression d'un gène est difficile du fait des localisations imprécises de ces régions eQTL ; elle nécessite *in fine* des expérimentations de biologie moléculaire lourdes à mettre en place.

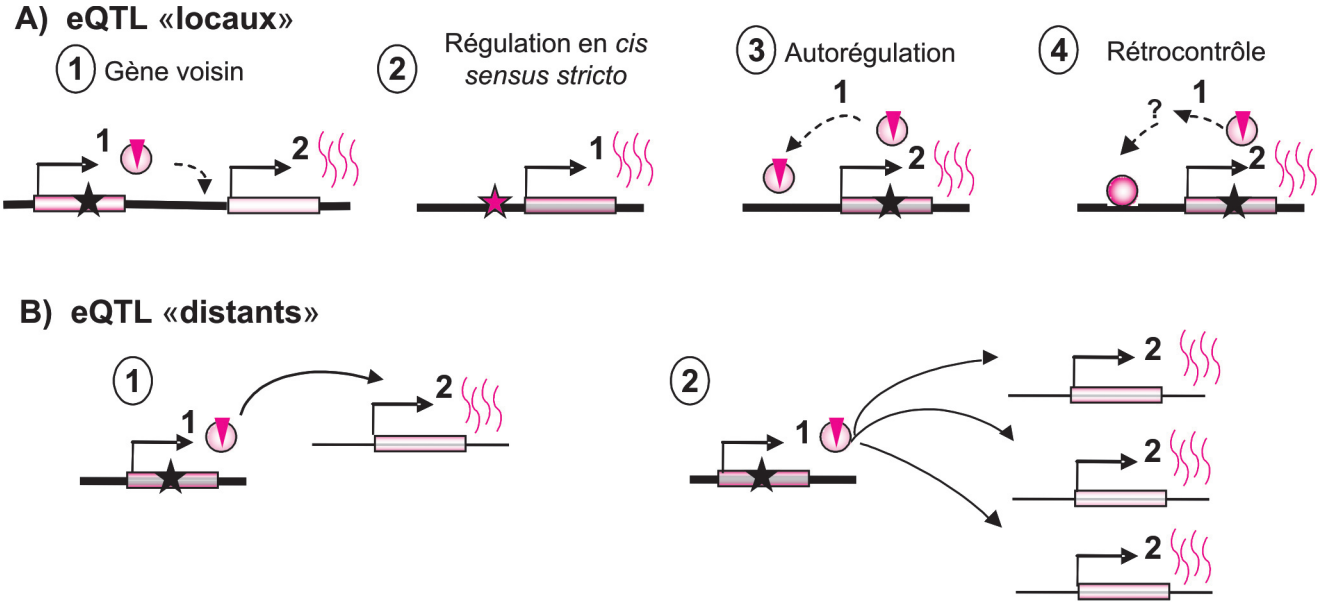
Les régions *trans* eQTL seraient quant à elles causées par des mutations affectant l'activité de gènes de la région qui réguleraient alors le niveau transcriptionnel d'autres gènes localisés ailleurs dans le génome (Farrall 2004). On peut donc imaginer des facteurs de transcription de toutes sortes. Néanmoins, différents auteurs ayant détecté plusieurs régions *trans* eQTL n'observent pas systématiquement la présence de gènes codant des facteurs de transcription dans ces régions (Yvert *et al* 2003, Bing et Hoeschele 2005). Il semblerait donc que l'intervention de gènes participant autrement à la machinerie transcriptionnelle (protéine d'épissage, protéine chaperonne, import/export nucléaire…) ou intervenant dans la dégradation des ARNm soit beaucoup plus fréquente que prévue.

1.3 / Les master eQTL

Les régions eQTL contrôlant l'expression d'un grand nombre de gènes sont communément appelées régions master eQTL (Morley *et al* 2004). Certains auteurs utilisent ce terme lorsque le nombre de gènes régulés par une telle région excède 25 (Gibson et Weir 2005).

A l'instar des régions *trans* eQTL, bien qu'on puisse imaginer que les master eQTL soient le reflet de l'action de gènes codant des facteurs de transcription, des analyses chez la levure avec une densité de marqueurs raisonnable permettent de conclure que la plupart de ces eQTL ne possèdent pas de gènes codant des facteurs de transcription, suggérant donc d'autres types de régulateurs (Yvert *et al* 2003).

Figure 2. *Régions eQTL contenant une mutation agissant localement (A) ou à distance (B).*



La région eQTL est indiquée par un trait gras (▬). Le gène porteur de la mutation est indiqué par une boite rouge (▭). La mutation est indiquée par une étoile rouge (★) quand elle est située dans les régions régulatrices du gène (région promotrice ici) provoquant alors une variation de quantité d'ARNm ( ) et par une étoile noire (★) quand elle est située dans la région codante provoquant une modification de l'activité de la protéine associée ( ).

2 / Cartographie de QTL d'expression : principaux résultats extraits de la littérature

Les études de la régulation génétique à l'échelle d'un ou quelques gènes existent depuis longtemps (Jacob et Monod 1961). En revanche, les nouvelles méthodologies, plus exhaustives et dites à haut débit, reposant en partie sur l'utilisation de puces à ADN n'ont réellement débuté qu'au début des années 2000. La technologie de puce à ADN s'est depuis démocratisée avec des coûts d'expérimentation (hors analyse) raisonnable de l'ordre de 150 euros par échantillon permettant d'analyser le profil transcriptomique de milliers de gènes d'un tissu pour des dizaines voire quelques centaines d'animaux. Par ailleurs, le développement de cartes génétiques relativement denses en marqueurs, disponibles pour la majorité des espèces modèles ou d'intérêt économique, ainsi que le savoir-faire des équipes qui ont dans un premier temps travaillé sur la cartographie de QTL, ont conduit à un essor des travaux de cartographie de QTL d'expression dont une revue est présentée ci-après.

2.1 / Proportion de gènes régulés par un QTL d'expression

a) Proportion de gènes ayant au moins un eQTL

Parmi les études de «génétique génomique» conduites essentiellement chez la levure, la souris et l'Homme, la proportion de gènes dont l'expression est détectée comme régulée par au moins un eQTL peut largement varier. Les différences sont dues à différents facteurs que sont la puissance de l'analyse reposant essentiellement sur l'effectif du dispositif d'animaux et le nombre de marqueurs utilisés ou encore le choix des seuils de signification d'un eQTL. Par exemple, seulement 0,6% des 374 gènes analysés dans l'étude de Stranger *et al* (2005) présentent au moins un eQTL. Ce pourcentage, un des plus bas observé, est en partie dû au manque de puissance de l'analyse d'association qui a été effectuée sur seulement 60 hommes non apparentés ; à noter également que le seuil de signification a été corrigé pour les tests multiples par la méthode de Bonferroni, méthode la plus drastique. A l'opposé, 59% des 5700 gènes analysés dans l'étude de Brem et Kruglyak (2005) présentent au moins un eQTL, cette étude a été réalisée sur un dispositif plus puissant, à savoir par analyse de liaison effectuée sur une structure en ségrégation de 112 levures. Ce pourcentage est de 40% dans l'étude de Yvert *et al* (2003) fondée sur 86 levures et diminue à 9% dans l'étude de Brem *et al* (2002) n'utilisant plus que 40 levures, ces études utilisant des seuils de signification similaires pour détecter un eQTL. Comme l'illustrent ces exemples, la puissance des dispositifs est un élément majeur dans les différences de résultats observées. Dans la majorité des études qu'ont recencées Williams *et al* (2007), les proportions de gènes ayant au moins un eQTL sont de l'ordre de 10 à 30%.

b) Proportion de cis et trans eQTL

D'après le tableau 1, le pourcentage de gènes pour lesquels une région eQTL est considérée comme agissant en *cis* est très variable d'une étude à l'autre (0,8 à 98%). Ce pourcentage peut varier en fonction de la taille de la fenêtre séparant les positions du gène régulé et du marqueur (= fenêtre *cis*) ainsi que du niveau de signification fixé pour définir une liaison génétique et/ou une association entre marqueur et expression d'un gène. Selon les études, la définition de la taille des fenêtres *cis* varie de 10kb (Brem *et al* 2002) à 20Mb (Bystrykh *et al* 2005). Certains auteurs font même référence aux *cis* eQTL lorsque l'eQTL régulant le niveau d'expression du gène correspond au marqueur le plus proche de l'oligonucléotide (Chesler *et al* 2005, Emilsson *et al* 2008). Des éléments *cis* régulateurs sont connus pour agir à plus de 1Mb du gène qu'il régule (Pfeifer *et al* 1999) ; aussi des tailles des fenêtres qui peuvent parfois paraître démesurées ne sont pas invraisemblables. De façon logique, le pourcentage de gènes régulés en *cis* augmente lorsque la fenêtre *cis* est élargie. Par ailleurs, les régions *cis* eQTL ont un effet plus prononcé sur les variations des niveaux d'ARNm des gènes comparé aux régions *trans*, du fait probablement d'un nombre d'événements biologiques moins élevé entre le polymorphisme et son effet. Ces régions sont en conséquence identifiées avec des statistiques de test plus élevées. Pour illustrer ce point, les études de Schadt *et al* (2003) menées chez la souris identifient comme *cis* eQTL 34% des 3701 régions eQTL localisées avec un LOD score > 4,3 et 71% des 784 régions eQTL identifiées avec un LOD score > 7. La proportion de régions eQTL agissant en *cis* tend donc à augmenter lorsque le seuil de signification des eQTL augmente (Schadt *et al* 2003, Doss *et al* 2005).

c) Proportion de master-trans eQTL

Il existe dans la littérature différentes façons de définir les gènes dont le niveau d'expression est régulé par une même région eQTL. Plusieurs études comptent le nombre de transcrits qui sont cartographiés pour un même marqueur (Bystrykh *et al* 2005, Chesler *et al* 2005, Hubner *et al* 2005, Cotsapas *et al* 2006). D'autres études comptent le nombre de gènes ayant un eQTL cartographié dans une fenêtre de taille prédéfinie (Brem *et al* 2002, Schadt *et al* 2003, Yvert *et al* 2003, Morley *et al* 2004).

Pour certains dispositifs eQTL actuellement décrits dans les deux espèces Homme et souris (tableau 1), le nombre de régions *master* eQTL varie de 1 (Cotsapas *et al* 2006) à 17 (Bystrykh *et al* 2005). En revanche, Monks *et al* (2004) et Emilsson *et al* (2008) n'en détectent aucune et suggèrent alors que ces régions ne seraient pas universelles dans le règne animal.

2.2 / Proportion de gènes régulés par plusieurs QTL d'expression

La plupart des analyses eQTL effectuées aujourd'hui utilise des méthodes «simple *locus*» où chaque locus est analysé indépendamment des autres locus pour détecter des liaisons ou associations avec les données transcriptomiques. Des analyses de liaison multi-QTL ou des tests d'associations multiples sont encore peu utilisés du fait de la complexité des tests statistiques, de l'importance des ressources informatiques nécessaires pour les effectuer et également et surtout de la taille requise des dispositifs. Néanmoins, des transcrits régulés par plusieurs endroits du génome peuvent être identifiés (Brem *et al* 2002, Schadt *et al* 2003, Morley *et al* 2004, Monks *et al* 2004, Brem et Kruglyak 2005, Cheung *et al* 2005, Hubner *et al* 2005, Stranger *et al* 2005, Cotsapas *et al* 2006). Ces études montrent que seulement 3% des phénotypes d'expression seraient contrôlés par un seul locus alors que plus de 50% seraient sous l'influence d'au moins 5 régions eQTL. Il n'y aurait que 23% des transcrits qui seraient régulés par un eQTL expliquant plus de 50% de la variance génétique. Ces observations sont cohérentes avec les mécanismes multiples de régulation de la quantité des ARNm d'un gène (cf. § 1.1). Tout comme les caractères complexes visibles à l'échelle de l'animal, les quantités d'ARNm sont à juste raison considérées comme des caractères quantitatifs dont les variations reposent sur un modèle polygénique additif avec quelques QTL à effets moyens à forts.

2.3 / Les interactions épistasiques

L'épistasie entre locus peut se définir comme étant l'interaction entre deux locus (ou plus) avec pour conséquence

Tableau 1. *Pourcentage de cis eQTL observés dans différentes études.*

Etude	Espèce (effectif de la population analysée)	Tissu	Nombre de gènes analysés	% de *cis eQTL*	Taille fenêtre *cis*	Seuil de signification	Nombre de *master eQTL*
Brem *et al* (2002)	Levure (40)	-	6215	36	10kb	P < 5 ×10 - 5	8
Yvert *et al* (2003)	Levure (86)	-	6215	25	10kb	P < 3,4 ×10 - 5	13
Schadt *et al* (2003)	Souris (111)	Foie	7861	34	2cM	LOD > 4,3	7
Monk *et al* (2004)	Homme (167)	Lignées cellulaires lymphoblastoïdes	2430	39	5Mb	P < 5 × 10 - 5	0
Morley *et al* (2004)	Homme (195)	Lignées cellulaires lymphoblastoïdes	3554	22	5Mb	P < 4, 3×10 - 7	2
Hubner *et al* (2005)	Rat (22)	Rein	15923	32	10Mb	P < 0,05	2
Chesler *et al* (2005)	Souris (32)	Prosencéphale	608	94	eQTL= marqueur le plus proche de l'oligo cible	FDR = 0,05	7
Bystrykh *et al* (2005)	Souris (22)	Cellules souches hématopoiétiques	12422	13	20Mb	P ≤ 0,005	17
Lan *et al* (2006)	Souris (60)	Foie	45000	12	10cM	LOD > 3,4	15
Wang *et al* (2005)	Souris (312)	Foie	23574	31	20Mb	P < 5 × 10 - 5	7
Cotsapas *et al* (2007)	Souris (31)	Cellules adipeuses	-	29	5Mb	Bonferroni alpha = 0,05	1
		Cerveau	17706	0,8			2
		Rein	9237	5,5			2
		Foie	10728	3,4			3
Bhasin *et al* (2008)	Souris (203)	Macrophage	17632	20	20Mb	LOD > 3	11
Schadt *et al* (2008)	Homme (427)	foie	39280	87	1Mb	Bonferroni alpha = 0,05	-
Emilsson *et al* (2008)	Homme (1002)	Sang	23720	98	eQTL= marqueur le plus proche de l'oligo cible	FDR = 0,05	0
Ghazalpour *et al* (2008)	Souris (110)	Foie	24048	40	10Mb	FDR = 0,1	4
Ponsuksili *et al* (2008)	Porc (74)	*Longissimus dorsi*	23256	7	Colocalisation QTL	P ≤ 0,05	-

un effet sur un caractère. En terme qualitatif (aussi appelé Mendélien), les interactions entre deux locus vont aboutir à l'atténuation ou disparition des effets de certains allèles d'un des deux locus selon la présence d'allèles à l'autre locus. En terme quantitatif, l'épistasie se réfère à la part de la variance génétique qui ne peut être expliquée ni par les effets additifs des allèles en présence, ni par les effets de dominance. L'effet de la coségrégation de plusieurs marqueurs sur la variabilité d'expression d'un gène a été testé chez la levure (Storey *et al* 2005). Ces travaux montrent que les niveaux de 37% des transcrits seraient régulés par au moins deux locus dont 14% seraient sous le contrôle de régulations épistasiques. Chez la drosophile, Anholt *et al* (2003) démontrent également l'importance des interactions épistasiques entre eQTL. Ces interactions entre eQTL régulant un même transcrit ne sont pas surprenantes au regard des connaissances que l'on a sur la régulation génique. En effet, la régulation des transcrits est généralement due à l'action de protéines agissant en combinaison. Ces premières études restent à être enrichies par de futurs travaux réalisés sur des dispositifs avec des effectifs plus élevés permettant de gagner en puissance. Par ailleurs, les algorithmes permettant de tester l'épistasie entre eQTL font l'objet de recherches importantes en particulier pour les adapter au nombre très élevé de variables à étudier (dizaines de milliers de gènes) (Carlborg *et al* 2005).

En conclusion de cette partie 2, les études évoquées conduisent à des résultats parfois disparates concernant les nombres d'eQTL de type *cis*, d'eQTL recencés par gène, d'eQTL en interaction. Ces disparités résultent probablement de la disparité des dispositifs expérimentaux analysés et des méthodes d'analyse utilisées. On peut ainsi recencer entre les études des différences dans le nombre d'individus analysés, la complexité génétique de la population en ségrégation étudiée (population F2 issus du croisement entre lignées consanguines ou non), les différentes sources biologiques d'où sont extraits les ARNm (tissus, cellules), les méthodes de quantification des ARNm à base de puces à ADN plus ou moins exhaustives en terme de gènes déposés ou de qualité de fabrication, le type et le nombre de marqueurs génétiques (microsatellites *vs* SNP, 100 *vs* 100 000 marqueurs analysés), les «fenêtres» sur le chromosome prises en compte pour la définition des *cis/trans*-eQTL, les

Tableau 2. *Principales études ayant, par une approche eQTL, identifié un gène candidat sous-jacent à un QTL responsable d'un caractère d'intérêt.*

Etude	Echantillon	Source des marqueurs génétiques	Technologie d'hybridation utilisée	Gène candidat identifié	Caractère ou maladie	Méthode d'identification	Mutation causale
Schadt *et al* (2003) (papier princeps)	111 souris F2 (B×D). Foie	Microsatellite. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	dolichyl-diospho-oligosaccharide-protein glycosyltransferase	Obésité	cis-eQTL	?
				cation-transporting ATPase	Obésité	cis-eQTL	?
Hubner *et al* (2005)	22 rats RI B×H/H×B. Rein et cellules adipeuses	Puce Affymetrix rat. 15923 oligos	1011 marqueurs microsatellites autosomaux provenant de WebQTL	73 gènes à tester dans populations humaines	Hypertension	cis-eQTL + co-localisation QTL	?
Schadt *et al.* (2005)	111 souris F2 (B×D). Foie	Microsatellite. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	Hsd11b1	Obésité	LCMS+Knockout	?
Mehrabian *et al* (2005)	111 souris F2 (B×D). Foie	Microsatellites. 1 marqueur tous les 13 cM	Puce Agilent bicouleur. 23574 oligos.	Alox5	Obésité	cis-eQTL + co-localisation QTL+ Knockout	?
Yaguchi *et al* (2005)	113 souris F2 (B×D). Foie et tissus adipeux	Microsatellite. 227 couvrant le génome	RT-PCRq pour 76 parmi 106 gènes d'une région QTL	17 gènes candidats	Diabète	cis/trans-eQTL + co-localisation QTL	?
Mootha *et al* (2006)	54 hommes d'âge similaires avec des degrés de tolérance au glucose différents	-	Puce Affymetrix HG-U133A. 39000 transcripts.	PGC-1α	Diabète type 2	GSEA	? (Mutations non sens reportées dans 2 autres études)
Lum *et al* (2006)	300 souris F2 (B×D). Brain	SNP. 1200 couvrant le génome	Puce Agilent bicouleur. 23574 oligos.	Pttg1	Obésité/Diabète	cis-eQTL	?
Meng *et al* (2007)	111 souris F2 (B×D) 334 souris F2 (B×H)		RT-PCRq	Abbc6	calcification cardiaque dystrophique	cis-eQTL+Knockout	Délétion de 10 pb en 3'UTR du gène
Bao *et al* (2007)	42 souris RIL (B×D). Brain	SNP. 3795 couvrant le génome	Puce Affymetrix M430. 39000 transcripts	Adcy2	Résistance douleur induite par la chaleur (nociception)	trans-eQTL + co-localisation QTL	? (2 mutations non sens dans la séquence codante (effets non validés))
				Myo7a	Trouble neurologiques et comportementaux	cis-eQTL + co-localisation QTL	?
				Ttc8	Syndrome Bardet-Biedl	cis-eQTL + co-localisation QTL	?
				Ank2	Troubles activités locomotrice	cis-eQTL + co-localisation QTL+ Knockout (dans autre design)	?
Schadt *et al* (2008)	427 hommes caucasiens. Foie	SNP. 782476 couvrant le génome	Puce Agilent bicouleur. 39280 oligos.	Rps26	Diabète type I	eQTL + GWAS	?
				Sort1	Maladie artères coronariennes	eQTL + GWAS	?
				Celsr2	Concentration LDL cholestérol dans le sang	eQTL + GWAS	?
Ponsuksili *et al* (2008)	74 porc F2. Muscle	Microsattelite. 116 couvrant le génome	Puce Affymetrix porcine. 23937 oligos.	Ahnak	Capacité de rétention en eau du muscle	cis-eQTL	?

méthodes d'estimation de liaison génétique ou d'association, le choix des seuils de signification, la prise en compte des tests multiples. Malgré ces différences, les approches combinant des données génétiques et génomiques permettent un nouvel angle d'étude de la variabilité d'expression des gènes et de leurs régulations que l'on sait complexes et aujourd'hui très partiellement connues. Notons que les méthodes permettant d'observer des modules géniques (présentées dans le § 3.3) peuvent également contribuer à cette connaissance. Enfin la prise en compte des effets épistasiques entre eQTL est également un élément important dans le décryptage de ces régulations.

3 / Apport des données transcriptomiques à la cartographie de QTL – Etat des lieux

Dans le cadre de l'identification de QTL responsables de la variation de caractères complexes, ces approches de génétique génomique offrent également une opportunité nouvelle pour caractériser ces régions QTL et faciliter l'identification des gènes causaux.

Les études utilisant les données transcriptomiques pour mieux caractériser des QTL contrôlant un caractère complexe sont en augmentation constante. La majorité de ces études, recencées dans le tableau 2, consiste à identifier des eQTL co-localisant avec la région QTL d'intérêt, donnant ainsi des informations fonctionnelles sur la région. Celles-ci peuvent dans certains cas permettre d'identifier le gène candidat positionnel et fonctionnel recherché. Par ailleurs, quelques auteurs ont utilisé les données transcriptomiques pour décomposer le caractère d'intérêt grâce aux nombreux phénotypes élémentaires que sont les niveaux de transcrits. Nous commencerons donc par exposer cette approche pour ensuite aborder l'approche eQTL plus couramment utilisée. Nous aborderons enfin différentes variantes visant à cartographier plus finement une région QTL.

3.1 / Décomposition d'un caractère complexe en phénotypes plus élémentaires

Cette stratégie consiste à identifier dans la population où des QTL ont été détectés, des sous-groupes de descendants partageant des profils transcriptomiques similaires. L'hypothèse alors faite est que ces sous-groupes d'individus homogènes transcriptionnellement le sont également génétiquement, conséquence possible de différentes mutations dont celles influençant le caractère d'intérêt. C'est d'autant plus vrai lorsque ces sous-groupes correspondent à des sous-groupes du caractère d'intérêt révélant ainsi la multiplicité des processus biologiques et donc des déterminants génétiques pouvant conduire à un même phénotype : par exemple des individus «gras» pourraient se diviser en deux sous-groupes, l'un caractérisé par la présence d'allèles codant des enzymes peu efficaces pour brûler les graisses, l'autre caractérisé par la présence d'allèles codant des enzymes très efficaces pour synthétiser des graisses à partir des sucres alimentaires. Une analyse de liaison sur le caractère d'intérêt en utilisant une partie de ces sous-groupes, peut alors révéler de nouveaux QTL, non observés de façon significative sur l'ensemble du dispositif, du fait de l'hétérogénéité du déterminisme génétique dans l'échantillon d'origine. Cette hétérogénéité peut être due aux influences polygéniques, aux interactions entre des QTL, ou peut-être encore aux effets environnementaux sur certains gènes.

Schadt *et al* (2003), ont été les premiers à développer ce concept. Ils l'ont appliqué à un dispositif de souris F2 issues d'un croisement de deux lignées consanguines en vue de mieux caractériser les QTL responsables du poids de tissu adipeux. Les transcriptomes hépatiques correspondant à 23 574 gènes ont été analysés. Après avoir identifié 208 transcrits corrélés au caractère et conservé une quarantaine de souris ayant les valeurs de caractère les plus extrêmes, les auteurs ont alors effectué une classification ascendante hiérarchique permettant de classer les individus selon leur profil transcriptome. Cette classification a permis de discriminer les animaux gras des maigres mais aussi de distinguer au sein de ces deux groupes d'extrêmes deux sous-groupes d'individus gras et maigres. Une nouvelle analyse QTL réalisée sur ces sous-groupes pris séparément a permis à la fois de confirmer de façon plus significative un QTL mais également de détecter un autre QTL non observé sur la population entière d'animaux (Schadt *et al* 2003).

Ce concept a été repris et appliqué à une famille de 50 poulets de chair, descendants d'un père connu pour être hétérozygote à un QTL pour le poids de tissu adipeux abdominal, localisé sur le chromosome 5 à environ 170 cM. Une analyse des corrélations prenant en compte la dépendance entre gènes, a permis d'identifier 688 transcrits corrélés au caractère «poids de gras» (Blum *et al* 2010). Une double classification sur ces gènes et individus (figure 3) met en évidence cinq groupes d'animaux présentant des profils transcriptomiques différents, en particulier deux sous-goupes pour chacun des groupes extrêmes gras et maigres. Une nouvelle analyse QTL après exclusion des huit animaux du sous-groupe n°5, un sous-groupe d'individus maigres, permet de détecter un autre QTL significatif à environ 100 cM, non observé sur la population entière d'animaux (figure 3). Une analyse fine des haplotypes des huit individus de ce sous-groupe montre qu'ils possèdent tous l'haplotype «q» au QTL à 170cM, suggérant une interaction entre les deux QTL, interaction qui a depuis été démontrée.

Ces études menées chez la poule et chez la souris, démontrent l'intérêt de cette approche pour mieux caractériser les QTL responsables d'un caractère complexe, et ce dans deux dispositifs d'animaux d'effectif réduit (environ 50 animaux) et de structure génétique assez différente.

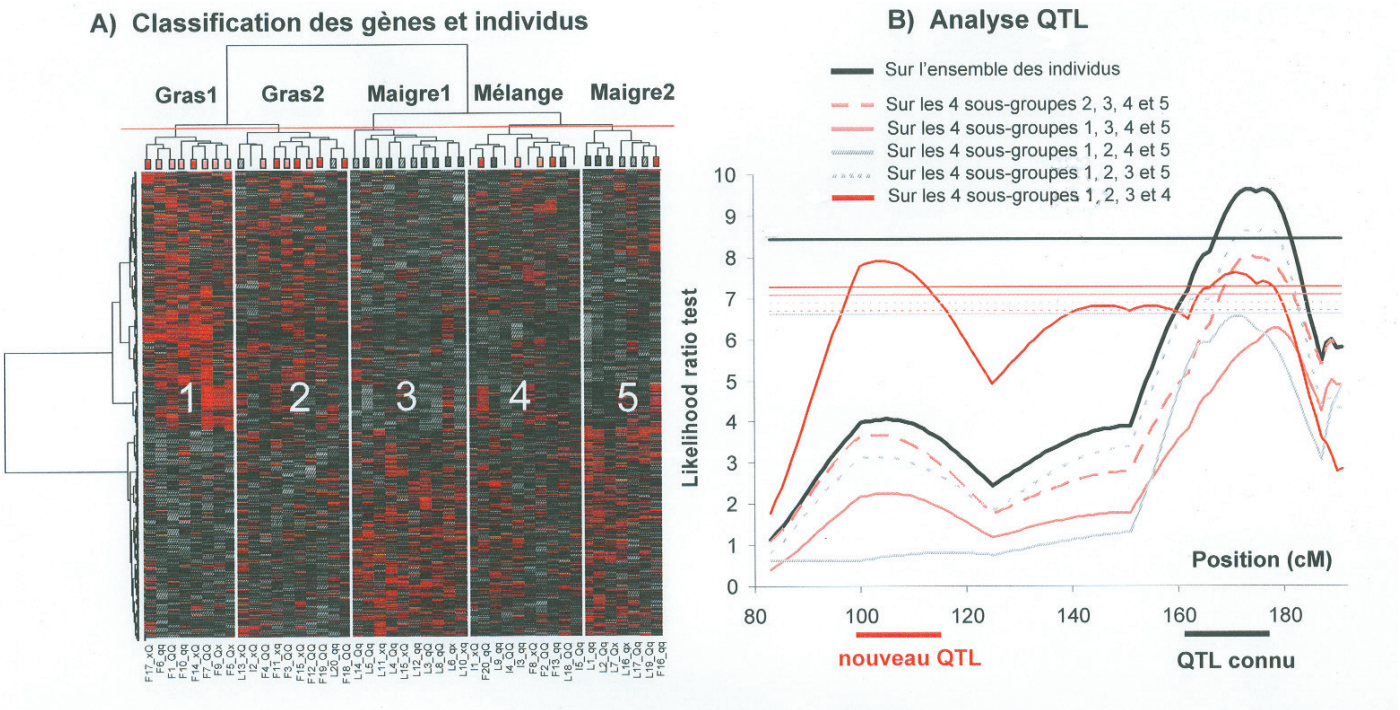
3.2 / Identification de régions eQTL co-localisant avec les régions QTL d'intérêt

L'objectif final de la cartographie de régions QTL est d'identifier le ou les gènes responsables de la variabilité d'un caractère quantitatif. Néanmoins, comme déjà rappelé dans l'introduction, les régions QTL comprennent pour la plupart des centaines de gènes candidats positionnels. La cartographie de régions eQTL a donc pour objectif de faciliter l'identification des meilleurs gènes candidats positionnels en apportant une information fonctionnelle les concernant. Cette approche peut également aider à affiner la localisation de régions QTL préalablement détectées. Cependant, on attend beaucoup plus des méthodologies de génotypage haut débit et de reséquençage concernant ce dernier point.

a) Principes et limites

Le principe est d'identifier les régions eQTL qui colocalisent avec les régions QTL d'intérêt (que nous appelerons régions eQTL/QTL), les gènes régulés pouvant eux se trouver n'importe où dans le génome. L'idée est alors de considérer que parmi les gènes régulés par la région, il y a ceux qui le sont par la mutation causale recherchée, apportant ainsi des informations fonctionnelles sur la mutation, de par les gènes qu'elle régule. Néanmoins l'identification de tels gènes, intermédiaires entre la mutation recherchée et le caractère d'intérêt, n'est pas si simple car, comme indiqué dans la figure 4, une région eQTL/QTL peut recouvrir bien d'autres

Figure 3. *Approche «décomposition d'un phénotype complexe en phénotypes plus élémentaires».*



(A) Classification des gènes et individus : Les individus en colonnes ont été classés selon leur profil transcriptomique sur la base des 688 gènes (situés en lignes) ayant une expression corrélée au caractère «poids de gras». Les individus en rouge et noir (plus ou moins intense selon le poids de gras) correspondent aux 20 individus extrêmes gras et maigres respectivement (F11 à F20 et L11 à L20).

(B) analyse QTL du poids de gras sur le chromosome 5, en utilisant l'ensemble des individus (courbe noire) ou en enlevant un des 5 sous-groupes observés après classification (autres couleurs). Dans le premier cas un QTL est observé à droite du chromosome 5 ; l'élimination des individus du sous-groupe 5 fait apparaître un QTL à gauche du chromosome 5 (courbe rouge). Les distances génétiques (cM) et la statistique de test de présence d'un QTL à une position donnée (LRT) sont indiquées sur les axes X et Y respectivement.

relations entre gènes régulés et caractère d'intérêt. On peut distinguer 4 types de relations :

– l'expression d'un ou plusieurs gènes contrôlée par la région eQTL/QTL peut être régulée par une mutation proche de la mutation contrôlant le caractère d'intérêt (cas 1 de la figure 4). En corollaire, plus ces deux mutations sont proches, plus elles sont en déséquilibre de liaison et plus les deux caractères «gène régulé» et «caractère d'intérêt» sont corrélés (Georges 2007). En conséquence l'identification de tels gènes apporte, par leurs corrélations avec le caractère, une information nouvelle sur la position la plus vraisemblable du QTL et devrait ainsi permettre par des méthodes multivariées de cartographie (Gilbert et Le Roy 2003) de réduire son intervalle de localisation. Cependant, de tels gènes n'apportent aucune information fonctionnelle sur la mutation causale recherchée ;

– l'alternative à la situation précédente est que l'expression d'un ou plusieurs des gènes contrôlés par la région eQTL/QTL soit régulée par la même mutation que celle contrôlant le caractère d'intérêt. Comme dans le cas précédent, de tels gènes nous apportent une information de position sur la mutation causale du QTL mais apportent aussi une information fonctionnelle sur les perturbations biologiques provoquées par cette mutation. Une telle information peut ainsi conduire à la mise en évidence d'un gène candidat positionnel et fonctionnel comme responsable du caractère d'intérêt. Ce cas idéal dans notre contexte correspond au cas 2 de la figure 4 mais ce n'est malheureusement pas la seule situation à envisager dans le cadre de gènes et caractère régulés par une même mutation ;

– la mutation causale au QTL peut contrôler les niveaux d'ARNm de gènes impliqués dans d'autres processus biologiques que ceux responsables de l'établissement du caractère complexe (cas 3 de la figure 4) ;

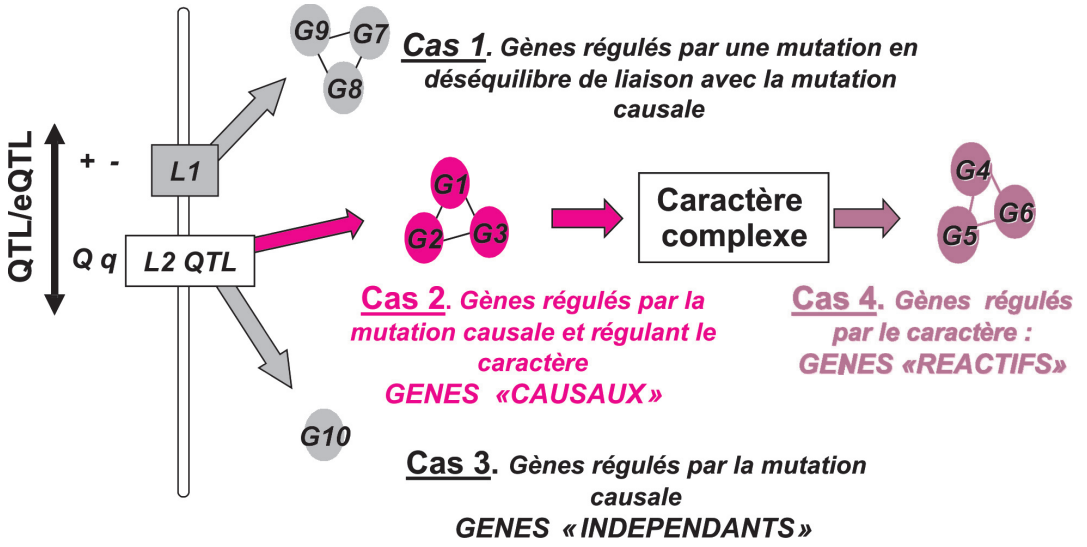
– enfin le caractère peut également réguler *a posteriori* le niveau d'autres ARNm (cas 4 de la figure 4).

Ces trois dernières situations faisant intervenir différentes relations entre caractère d'intérêt et gènes régulés par une région eQTL/QTL ont conduit certains auteurs à adopter une nomenclature pour les qualifier. Schadt *et al* (2005) ont ainsi introduit les termes de gènes «causaux», «indépendants» et «réactifs» par rapport au caractère, termes repris dans la figure 4**.**

Une spécificité de ces gènes «causaux», «réactifs» et «indépendants» est qu'ils sont tous régulés par la même région eQTL/QTL d'intérêt compliquant l'identification des gènes «causaux» qui sont les seuls donnant des informations fonctionnelles directes sur le gène et la mutation causale sous-jacente au QTL d'intérêt. Généralement, les chercheurs procèdent à une analyse plus fine des fonctions des gènes régulés par la région eQTL/QTL en espérant identifier des gènes dont la fonction biologique ait un lien évident avec le caractère d'intérêt. Ces gènes sont alors considérés comme des gènes «causaux» et peuvent ouvrir des pistes quant au meilleur gène candidat positionnel et fonctionnel de la région QTL qui à la fois contrôle leur expression et contrôle le caractère d'intérêt.

Concernant les espèces d'élevage, quelques études ont mis en œuvre cette approche et ont ainsi identifié un ou des gènes en lien avec le caractère d'intérêt et régulé(s) par la région QTL, en faisant donc une bonne signature fonctionnelle du gène candidat positionnel et fonctionnel recherché. On peut ainsi citer une étude conduite par Ponsuksili *et al* (2008) sur le pouvoir de rétention d'eau de la viande chez le porc, une autre sur le poids de gras chez le poulet de chair (Le Mignon *et al* 2009, Blum *et al* 2010) ou encore sur la couleur de la

Figure 4. *Différentes relations pouvant exister entre le caractère d'intérêt et les gènes «contrôlés» par une région QTL/eQTL.*



Quatre types de relations entre «région eQTL/QTL», «gènes» et «caractère» peuvent être distingués :

Cas 1 : l'expression d'un ou plusieurs gènes est contrôlée par la région eQTL/QTL par une mutation proche de la mutation causale recherchée contrôlant le caractère d'intérêt.

Cas 2, 3 et 4 : l'expression d'un ou plusieurs gènes est contrôlée par la région eQTL/QTL par la même mutation que celle contrôlant le caractère. Cependant dans le **cas 2**, les gènes sont responsables d'une part de la variation du caractère. Dans le **cas 3**, ils sont régulés par les variations du caractère (reflet bien souvent de boucles de rétrocontrôle). Enfin, dans le **cas 4**, les gènes sont contrôlés par la mutation causale indépendamment du caractère d'intérêt.

viande chez le poulet de chair (LeBihan-Duval *et al* communication personnelle). La dernière étude est la seule à notre connaissance qui ait abouti à l'identification des mutations causales au QTL d'intérêt ; cette étude correspond à la situation favorable des *cis* eQTL (voir ci-après).

b) Cas particulier des gènes régulés par un cis eQTL

Les gènes dont le niveau d'ARNm est régulé par une région eQTL de type *cis* co-localisant avec une région QTL sont des gènes particulièrement intéressants.

Tout d'abord, comme déjà indiqué précédemment, les régions *cis* eQTL sont plus facilement détectables que les régions *trans* eQTL (LOD score élevé) probablement dû à un nombre réduit d'événements biologiques séparant la mutation dans un gène et sa propre régulation transcriptionnelle. Afin de mettre en évidence expérimentalement cette caractéristique donnée aux régions *cis* eQTL, Schadt *et al* (2003) ont mené un programme de cartographie d'eQTL sur un croisement de lignées de souris DBA et B6 connues pour différer par une délétion de 2 paires de bases dans le gène C5, cette délétion affectant la dégradation de ses transcrits. Comme attendu, ils ont identifié au niveau du gène C5 un *cis*-eQTL avec un LOD score hautement significatif, et même un des plus élevés de l'étude (LOD score > 27,4).

Par ailleurs, l'analyse de la fonction de ces *cis* eQTL permet, si cette fonction est en relation avec le caractère d'intérêt, d'émettre une hypothèse simple sur le meilleur gène candidat positionnel et fonctionnel de la région QTL d'intérêt ainsi que sur la position de la mutation causale dans le gène candidat : ainsi le gène recherché est le gène *cis* eQTL qui est alors responsable du caractère et régulé par la mutation eQTL/QTL située dans ses régions régulatrices.

Prenons l'exemple des travaux de Le Bihan-Duval *et al* (communication personnelle) dans lesquels moins de deux années se sont écoulées entre la primo-localisation d'une région QTL contrôlant la couleur de la viande (Nadaf *et al* 2007) et la détection de la mutation causale associée, montrant ainsi l'efficacité de l'approche. Une analyse des gènes présents dans la région QTL a révélé l'existence d'un gène, *BCMO1*, qui code une enzyme clef de la dégradation du β-carotène, pigment dont l'effet sur la couleur des tissus est bien connu. Ce gène représente donc un bon candidat à la fois positionnel et fonctionnel pour le QTL d'intérêt. Par ailleurs, l'expression de ce gène est contrôlée par une région eQTL qui co-localise avec la région QTL affectant la couleur de la viande. Ces résultats suggéraient donc que *BCMO1* était le gène causal et que la mutation causale recherchée se trouvait dans les parties régulatrices du gène *BCMO1* ; cette mutation contrôlait ainsi la variation de l'expression de *BCMO1*, cette variation impactant alors la couleur de la viande. Le promoteur de ce gène a donc été séquencé sur des animaux de génotypes variés au QTL. Deux mutations dans le promoteur du gène *BCMO1* ont été identifiées et des expérimentations de biologie moléculaire ont montré que ces mutations avaient bien un effet sur l'expression du gène. Notons que la différence d'expression musculaire du gène *BCMO1* entre les deux génotypes au QTL est de deux écart-types contre seulement un écart-type pour celle de la couleur de la viande, illustrant ainsi que la différence entre génotypes d'un phénotype élémentaire (ici l'expression d'un gène) est supérieure à celle d'un phénotype plus complexe, ce qui rend les analyses de cartographie de eQTL/QTL plus puissante et plus précise.

Ainsi, après l'identification de gènes régulés par des régions eQTL co-localisant avec un QTL responsable d'un caractère complexe, la majorité des études se focalise sur les gènes ayant un eQTL de type *cis* (Binget Hoeschele 2005, Doss *et al* 2005, Yamashita *et al* 2005, GuhaThakurta *et al* 2006, Lum *et al* 2006, Ponsuksili *et al* 2008). Cependant, contrairement à l'exemple évoqué plus haut, peu d'études ont abouti à la découverte de la mutation causale, pour différentes raisons : *i)* des gènes présents dans la région eQTL/QTL (dont le gène causal régulé en *cis* recherché) peuvent avoir une

fonction partiellement connue voire même inconnue, ne permettant donc pas de faire le lien avec le caractère d'intérêt ; un grand nombre de gènes sont encore dans ce cas. De plus, tous les gènes de l'intervalle peuvent ne pas avoir été déposés sur la puce et ne sont par conséquent pas analysables. *ii)* la démonstration que le gène régulé en *cis* est le gène réellement responsable de la variabilité du caractère complexe n'est pas des plus simples. Une telle démonstration nécessite des expérimentations supplémentaires de biologie moléculaire qui sont assez longues à mettre en œuvre ; à défaut il est nécessaire que la fonction du gène soit clairement décrite comme étant en lien étroit avec le caractère d'intérêt (cf. point précédent). *iii)* Le nombre de gènes candidats régulés en *cis* présents dans l'intervalle de localisation du QTL peut parfois être élevé. *iv)* des régions *cis* eQTL peuvent être des faux positifs. En effet, Alberts *et al* (2005) ont observé que des polymorphismes de séquences dans la région codante de certains gènes peuvent plus ou moins influencer la qualité de fixation de l'ARNm cible à la sonde déposée sur la puce. Cette différence de fixation de l'ARNm selon les polymorphismes de séquence est ensuite interprétée à tort par l'expérimentateur comme une variation de la quantité d'ARNm. *v)* On peut aussi considérer la situation où des gènes régulés en *cis* dans la région QTL seraient également régulés par un autre endroit du génome (en *trans*) avec des phénomènes d'interaction, perturbant ainsi la détection du *cis* eQTL par des analyses classiques. Yaguchi *et al* (2005) proposent la construction de lignées congéniques qui seraient hétérozygotes uniquement pour la région eQTL/QTL d'intérêt privilégiant ainsi les interactions en *cis*. Cependant, la création d'une lignée congénique est souvent coûteuse et ne garantit pas de reproduire le phénotype assigné par le QTL.

Notons que la co-localisation eQTL/QTL peut également permettre d'exclure des gènes candidats fonctionnels. En effet, Lan *et al* (2004) ont identifié un *cis* eQTL très significatif (LOD 30) pour un gène candidat fonctionnel au diabète (*Pdi*) rapidement exclu puisque ce dernier ne co-localisait pas avec les régions QTL identifiées en parallèle dans le même dispositif.

Quelles que soient les études, le nombre de gènes régulés par une région eQTL/QTL est élevé. Différents facteurs concourent à ces observations : d'une part, les régions QTL ne sont pas localisées avec précision ; d'autre part, les études transcriptomiques permettent l'analyse de dizaines de milliers de gènes dont les processus de régulations sont complexes ; enfin le polymorphisme dans les dispositifs analysés (en général une population issue du croisement entre lignées divergentes) peut être elevé, suggérant de nombreuses mutations en déséquilibre de liaison dans une région eQTL/QTL. Afin de sélectionner les gènes les plus en lien fonctionnellement avec le caractère parmi ceux, parfois nombreux, présents dans une région eQTL/QTL, une analyse bibliographique peut etre menée. Néanmoins, celle-ci se révèle la plupart du temps très chronophage et parfois non informative, les connaissances sur la fonction des gènes étant partielles.

Aussi, des méthodes se sont développées pour diriger plus rapidement l'expérimentateur vers la mutation causale. Certaines n'utilisent d'ailleurs aucune information extérieure au jeu de données (génotype et transcriptome) et vise à distinguer les modules géniques de type «causaux» de ceux considérés comme «réactifs» ou «indépendants», les premiers étant particulièrement utiles dans notre contexte d'étude. D'autres méthodes au contraire utilisent des bases de données associant des informations fonctionnelles aux différents gènes d'une espèce et ont pour objectif d'identifier des modules géniques partageant une même fonction. De telles méthodes permettent dans notre contexte de sélectionner le module dont la fonction serait le plus en relation avec le caractère d'intérêt.

3.3 / Recherche de modules géniques

a) Sélection des modules géniques «causaux», «réactifs» et «indépendants»

Schadt *et al* proposent une méthode appelée LCMS pour *Likelihood-based Causality Model Selection* (Schadt *et al* 2005), permettant d'identifier la relation la plus probable qui existe entre un caractère complexe et l'expression d'un gène, tous deux contrôlés par une même région eQTL/QTL. La méthode considère les trois modèles présentés dans la figure 4 : le modèle «causal», le modèle «réactif» et le modèle «indépendant».

Pour chaque modèle, les paramètres sont estimés par maximisation du critère de vraisemblance et le modèle retenu correspond à celui minimisant le critère d'Akaike (critère couramment utilisé pour identifier le modèle le plus vraisemblable).

Ces auteurs ont appliqué leur méthode dans le cadre de l'identification des gènes causaux au gras viscéral dans un croisement de souris F2 (Schadt *et al* 2005). Après avoir identifié environ 100 gènes dont les niveaux d'ARNm sont à la fois corrélés au caractère et régulés par au moins deux régions eQTL co-localisant avec deux des 4 régions QTL contrôlant le caractère, les auteurs se sont concentrés sur les dix gènes ayant la plus forte probabilité d'être gènes «causaux». Ils ont ainsi pu identifier le gène *Hsd11b1* déjà connu pour son lien avec l'obésité (Masuzaki et Flier 2003). Drake *et al* (2006) ont poursuivi ces travaux en surexprimant ou réprimant chacun des 9 autres gènes chez des souris et ont étudié leur phénotype. Ils ont ainsi pu valider que 8 des 9 gènes ont un effet sur le caractère, deux d'entre eux ayant un effet opposé selon le sexe de l'animal. Ces gènes peuvent donc être considérés comme une signature fonctionnelle de la(des) mutations(s) des régions eQTL/QTL considérées.

Comme montré dans l'exemple précédent, les auteurs ont du procéder à des sélections de gènes pour obtenir les gènes «causaux» les plus vraisemblables. Sans sélection préalable, la méthode de LCMS peut prédire des milliers de gènes «causaux» dont beaucoup sont des faux positifs. En effet, des facteurs de variabilité inconnus ou non mesurés peuvent influer sur l'expression des gènes ainsi que sur la variabilité du caractère d'intérêt de sorte à faire apparaître un faux lien de causalité entre les deux (Kruglyak et Storey 2009). De plus, Schadt *et al* supposent qu'un gène corrélé au caractère peut être soit «causal», soit «réactif» et non les deux à la fois. Or, il se peut qu'il y ait des mécanismes complexes de rétrocontrôle et que les hypothèses posées soient donc trop restrictives. Néanmoins, l'exemple cité plus haut démontre l'intérêt de cette méthode.

b) Sélection de modules géniques partageant une même fonction

Ce paragraphe expose des méthodes recherchant à tirer profit de l'information des bases de données associant une fonction à un gène.

Concernant les bases de données d'information biologique, il en existe principalement deux : la base de données des termes Gene Ontology (GO) (Ashburner *et al* 2000) et la base de données des termes KEGG (Kanehisa *et al* 2006). Ces bases de données proposent d'associer à chaque gène les termes le caractérisant le mieux selon les connaissances du moment. A noter que les associations entre termes fonctionnels et gènes peuvent différer entre ces deux bases. La base de données des termes KEGG concerne en majorité des réactions biochimiques et donc des enzymes et les associations «termes

KEGG-gène» sont inférées manuellement. La base de données des termes GO, elle, est scindée en trois classes d'ontologie : les processus biologiques, les fonctions moléculaires et les composants cellulaires. Les associations «termes GO-gènes» sont pour la majorité inférées automatiquement par bioinformatique par exemple en prédisant la(les) fonction(s) d'un gène à partir de sa séquence par recherche de motifs fonctionnels. Aussi, les associations trouvées entre gènes et termes fonctionnels ne sont pas parfaites puisque parfois erronées pour certaines ou encore partielles par rapport à ce qui est connu dans la littérature. Malgré tout, ces informations permettent l'analyse fonctionnelle de nombreux gènes à la fois. Ces deux bases de données sont très largement utilisées et sont considérées comme complémentaires. La base KEGG est plus fiable mais moins complète que la base GO. Par ailleurs, il faut garder à l'esprit que la proportion de gènes ayant des fonctions inconnues ou partielles est importante ; en conséquence ces bases de données reflètent seulement la connaissance partielle que l'on a aujourd'hui de la fonction des gènes.

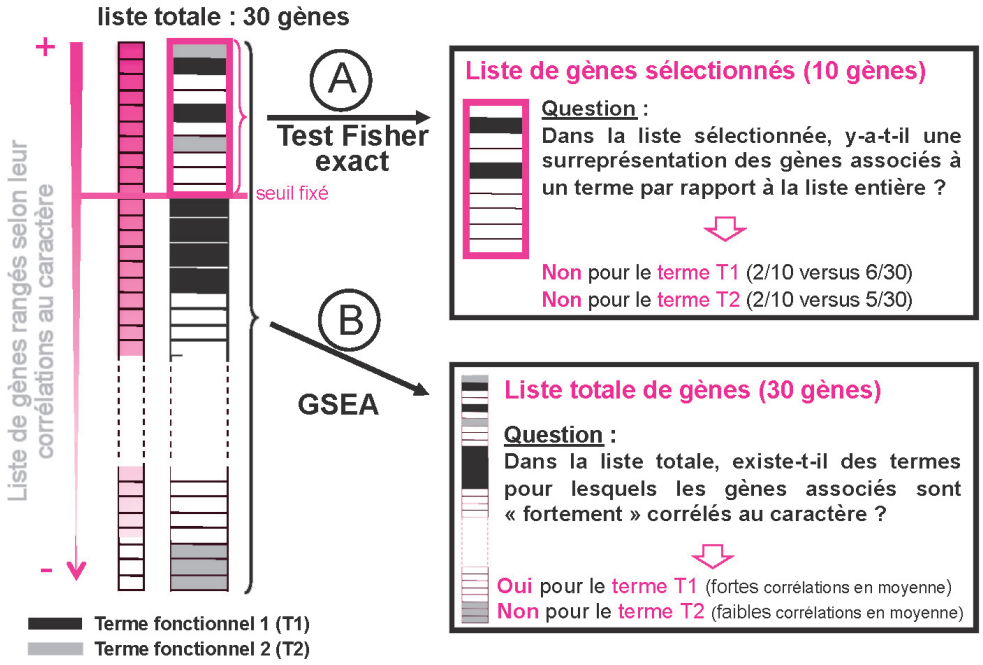
Concernant les méthodes permettant d'identifier des termes fonctionnels que ce soit des termes GO ou KEGG (et les gènes associés) en lien avec le caractère d'intérêt, il en existe deux couramment utilisées : il s'agit du test de Fisher exact et de la méthode GSEA pour *Gene Set Enrichment Analysis*, présentées ci-après et dans la figure 5.

Le test de Fisher exact consiste à mesurer, au sein d'une sous-liste de gènes d'intérêt, l'enrichissement de gènes associés à un terme fonctionnel particulier par rapport à la liste entière des gènes sur la puce. Autrement dit, il s'agit de tester la surreprésentation d'un terme fonctionnel dans une sous-liste de gènes. Ce test est effectué pour chaque terme fonctionnel permettant de recenser ceux qui caractérisent la sous-liste de gènes et qui sont donc impliqués dans la variabilité du caractère. En effet, la sous-liste de gènes d'intérêt est prédéfinie et correspond classiquement aux gènes corrélés au caractère ou encore aux gènes différentiellement exprimés entre individus extrêmes pour le caractère. Une des limites de cette méthode est de se focaliser seulement sur cette sous-liste, ce qui peut se révéler trop restrictif. Une fois la sous-liste d'intérêt sélectionnée, la méthode ne tient plus compte des valeurs de corrélation : tous les gènes sont considérés au même niveau alors que certains sont plus corrélés au caractère que d'autres. Par ailleurs, parmi les gènes non retenus, certains ont des corrélations en limite de signification (proche du seuil fixé) (figure 5).

Pour pallier ces problèmes, la méthode GSEA proposée par Mootha *et al* (2003), permet de trouver les termes fonctionnels en lien avec le caractère en considérant cette fois-ci la liste entière de gènes et leur corrélation avec le caractère (figure 5). D'un certain point de vue, il s'agit de réaliser une analyse différentielle non plus à l'échelle du gène mais à l'échelle du terme fonctionnel. Le principe de la méthode est de regarder au sein d'une liste de gène si un terme fonctionnel particulier est associé préférentiellement à des gènes fortement corrélés au caractère d'intérêt. Si c'est le cas, on pourra dire que ce terme est lié à la variabilité du caractère. Comme indiqué en figure 5, la méthode GSEA peut révéler des termes fonctionnels en lien avec le caractère d'intérêt non-identifiés par la méthode du Fisher exact lorsqu'ils sont associés à des gènes ayant une corrélation avec le caractère en limite de signification.

Différents logiciels permettent une mise en œuvre rapide de l'une ou l'autre de ces deux méthodes même si la date de mise à jour de la base de données d'annotations utilisée n'est pas toujours accessible. Les logiciels fondés sur le test du Fisher exact utilisent le plus souvent les termes GO : AmiGO (Carbon *et al* 2009), EasyGO (Zhou et Su 2007), GOtm (Zhang *et al* 2004) dont une version prend en compte les corrélations des gènes avec le caractère d'intérêt (GOrilla, Eden *et al* 2009), GOstats sous l'environnement R utilisant les termes GO et les termes KEGG (Falcon et Gentleman 2007). Quant à la méthode GSEA, elle est disponible sous l'environnement R avec le logiciel GSEA-P-R (Subramanian *et al* 2005), et plusieurs autres versions comme GSA dont certaines proposent leurs propres associations gènes-termes fonctionnels, les termes pouvant être la localisation chromosomique, le partage de motifs nucléotidiques (Kim et Volsky 2005, Efron et Tibshirani 2007, Jiang et

Figure 5. *Méthodes permettant de tester si un terme fonctionnel est associé significativement à un caractère complexe : en A, test Fisher exact, en B, méthode GSEA.*



Gentleman 2007, Subramanian *et al* 2007).

Les deux méthodes reposent donc sur la constitution de groupes de gènes selon leur appartenance à un terme fonctionnel, la méthode GSEA étant généralement plus puissante. Par ailleurs, selon la source d'information biologique utilisée, généralement la base de données des termes KEGG ou encore des termes GO, les associations entre termes fonctionnels et gènes peuvent parfois différer.

Dans le présent contexte de caractérisation de QTL, Ghazalpour *et al* (2005) ont utilisé ces deux méthodes afin de trouver les métabolismes liés au caractère gras subcutané dans un croisement de souris F2. Après avoir sélectionné les gènes différentiellement exprimés selon le poids en gras subcutané des individus, soit environ 5000 gènes, ils ont recherché les principaux métabolismes dans lesquels ces gènes interviennent. Pour cela, les auteurs ont appliqué à la fois la méthode GSEA et la méthode de Fisher exact en considérant comme liste entière de gènes la liste des 5000 gènes différentiellement exprimés et comme sous-liste de gènes les 20% les plus corrélés. Ces auteurs ont utilisé les termes KEGG pour constituer les groupes de gènes. Les deux méthodes se révèlent comparables, avec cependant une sensibilité légèrement supérieure pour la méthode GSEA : 13 métabolismes sont trouvés avec GSEA et 10 par le test de Fisher exact (tous inclus dans les 13). Les métabolismes trouvés correspondent à des métabolismes liés à l'énergie et aux lipides et concernent 150 gènes dont 68 sont corrélés au gras subcutané. Ce résultat rappelle le concept selon lequel la régulation d'un métabolisme n'implique le contrôle que d'une partie seulement des gènes qui y sont associés. Les auteurs ont alors voulu tester si ces 150 gènes étaient régulés par des régions communes du génome. Ils ont montré qu'il y avait une surreprésentation d'eQTL contrôlant ces gènes dans des régions QTL gras déjà identifiées (Drake *et al* 2001, Schadt *et al* 2003) apportant ainsi une information fonctionnelle sur ces dernières.

c) Cartographie fine d'un QTL par prédiction, sur la base d'une signature fonctionnelle, de l'allèle causal chez des animaux recombinants

Dans la continuité des paragraphes précédents, les niveaux d'expression des gènes identifiés comme signature fonctionnelle de la mutation causale d'intérêt (que nous appelerons gènes «signatures»), peuvent donc en théorie être utilisés comme prédicteurs de l'allèle à la mutation causale. Appliquée à des animaux recombinants, tous descendants d'un père par exemple, la prédiction de l'allèle paternel Q ou q reçu à la mutation causale, devrait permettre de réduire la région QTL d'intérêt. L'approche que nous avons développée en ce sens à l'UMR de génétique animale INRA Agrocampus Ouest, dans une famille d'une cinquantaine de descendants issus d'un père hétérozygote pour un QTL, peut être décomposée en 4 étapes :

– la première étape consiste à identifier, uniquement sur la base des marqueurs, les descendants ayant reçu la totalité de l'haplotype paternel Q (porteur de l'allèle causal Q) ou q (porteur de l'allèle causal q) de la région QTL (individus non recombinants) ;

– la seconde étape consiste alors à établir une fonction à partir du niveau d'expression des gènes «signatures» qui permet de discriminer au mieux l'haplotype Q de l'haplotype q ;

– cette fonction est ensuite utilisée pour prédire le statut de l'allèle causal (Q ou q) chez les animaux recombinants, sur la base du niveau d'expression des différents gènes «signatures» ;

– enfin en confrontant les fragments haplotypiques Q ou q transmis par le père avec l'allèle Q ou q prédit à la mutation causale des recombinants, l'intervalle de localisation du QTL peut être réduit. Le gain de réduction est d'autant plus important que le nombre de marqueurs dans la région est grand, permettant de déterminer avec précision le point de recombinaison chez chaque individu.

Nous avons ainsi réduit une région QTL responsable du poids de gras abdominal chez le poulet de chair de 31 cM à 7 cM (Le Mignon *et al* 2009). Cette approche peut donc se révéler très efficace sous réserve d'une signature fonctionnelle fiable de la mutation causale d'intérêt et de disposer d'animaux recombinants intéressants. A noter que pour les programmes ayant créé des recombinants maîtrisés, cette approche peut éviter de générer de nouveaux descendants pour identifier leur génotype au QTL, ce qui est particulièrement long et coûteux.

Conclusion

La génomique fonctionnelle vise à améliorer notre compréhension des fonctions et de la régulation de l'expression des gènes, de leurs transcrits et des protéines associées, à l'échelle globale du génome (pour revue, Hocquette *et al* 2009). Elle permet d'établir le pont entre le séquençage du génome et les phénotypes observés. Le développement de larges collections d'EST et surtout le séquençage de génomes complets couplé à la prédiction de séquences géniques a permis le développement de puces oligonucléotides composées de tout ou partie des gènes prédits ou connus d'un génome.

Toutes ces nouvelles ressources ouvrent des perspectives dans la compréhension des mécanismes de régulation transcriptionnelle, des réseaux de gènes ou bien des chemins métaboliques déterminant l'établissement d'un phénotype. La «génétique génomique», qui combine les données d'expression avec les données de polymorphisme génétique et qui recouvre d'autres approches que l'approche eQTL proprement dite, ouvre des perspectives intéressantes dans le cadre de l'identification des gènes responsables de la variabilité d'un caractère complexe. Il ne faut cependant pas perdre de vue ses limites, d'ordres économiques, techniques ou méthodologiques : *i)* La technologie reste onéreuse dès lors que des centaines d'animaux sont à analyser. *ii)* L'expression des gènes est mesurée dans un tissu précis à un temps précis, les résultats et les conclusions sont donc difficilement transposables et doivent ainsi être interprétés en conséquence. *iii)* La plupart des méthodes actuelles ne peuvent pas traiter les modèles non additifs, épistasiques ou résultant d'autres effets complexes sur la variabilité d'expression des gènes, de sorte que le nombre de gènes présentant des régions eQTL est probablement sous-estimé. *iv)* La plupart des analyses souffrent de puissance statistique limitée. Sans puissance statistique suffisante, les études se limitent aux gènes présentant de fortes variabilités d'expression. *v)* A l'heure actuelle, la proportion de gènes ayant des fonctions totalement inconnues ou partiellement connues est encore importante.

En revanche, la réduction des coûts des puces à ADN (puces qui seront remplacées dans un avenir plus ou moins proche par du séquençage haut débit), couplée au développement de puces à SNP à haute densité permettra d'améliorer dans un futur proche la puissance et la précision des analyses eQTL. Par ailleurs, les puces à SNP à très haut débit couplées à la possibilité maintenant de re-séquencer plusieurs individus devraient permettre de localiser très finement les régions QTL et d'en identifier tous les polymorphismes que l'on sait nombreux. Aussi, apporter de l'information fonctionnelle sur l'impact de la mutation causale dans une région QTL d'intérêt grâce aux approches de génétique génomique peut être un élément déterminant pour identifier cette mutation parmi les différents polymorphismes observés dans une région.

Références

Alberts R., Terpstra P., Bystrykh L.V., de Haan G., Jansen R.C., 2005. A statistical multiprobe model for analyzing *cis* and *trans* genes in genetical genomics experiments with short-oligonucleotide arrays. Genetics, 171, 1437-1439.

Anholt R.R., Dilda C.L., Chang S., Fanara J.J., Kulkarni N.H., Ganguly I., Rollmann S.M., Kamdar K.P., Mackay T.F., 2003. The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome. Nat. Genet., 35, 180-184.

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25-29.

Bing N., Hoeschele I., 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics, 170, 533-542.

Blum Y., Le Mignon G., Lagarrigue S., Causeur D., 2010. A factor model to analyze heterogeneity in gene expression. BMC Bioinformat., 11, 368.

Brem R.B., Kruglyak L., 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc. Natl. Acad. Sci., 102, 1572-1577.

Brem R.B., Yvert G., Clinton R., Kruglyak L., 2002. Genetic dissection of transcriptional regulation in budding yeast. Science, 296, 752-755.

Bystrykh L., Weersing E., Dontje B., Sutton S., Pletcher M.T., Wiltshire T., Su A.I., Vellenga E., Wang J., Manly K.F., 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using «genetical genomics». Nat. Genet., 37, 225-232.

Carbon S., Ireland A., Mungall C.J., Shu S., Marshall B., Lewis S., 2009. AmiGO: online access to ontology and annotation data. Bioinformatics, 25, 288-289.

Carlborg O., De Koning D.J., Manly K.F., Chesler E., Williams R.W., Haley C.S., 2005. Methodological aspects of the genetic dissection of gene expression. Bioinformatics, 21, 2383-2393.

Chesler E.J., Lu L., Shou S., Qu Y., Gu J., Wang J., Hsu H.C., Mountz J.D., Baldwin N.E., Langston M.A., 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet., 37, 233-242.

Cheung V.G., Spielman R.S., Ewens K.G., Weber T.M., Morley M., Burdick J.T., 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature, 437, 1365-1369.

Cotsapas C.J., Williams R.B., Pulvers J.N., Nott D.J., Chan E.K., Cowley M.J., Little P.F., 2006. Genetic dissection of gene regulation in multiple mouse tissues. Mamm. Genome, 17, 490-495.

De Vienne D., Leonardi A., Damerval C., Zivy M., 1999. Genetics of proteome variation for QTL characterization: application to drought-stress responses in maize. J. Exp. Bot., 50, 303-309.

DeCook R., Lall S., Nettleton D., Howell S.H., 2006. Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics, 172, 1155-1164.

Doss S., Schadt E.E., Drake T.A., Lusis A.J., 2005. *Cis*-acting expression Quantitative Trait Loci in mice. Genome Res., 15, 681-691.

Drake T.A., Schadt E., Hannani K., Kabo J.M., Krass K., Colinayo V., Greaser L.E., Goldin J., Lusis A.J., 2001. Genetic loci determining bone density in mice with diet-induced atherosclerosis. Physiol. Genomics, 5, 205-215.

Drake T.A., Schadt E.E., Lusis A.J., 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. Mamm. Genome, 17, 466-479.

Eden E., Navon R., Steinfeld I., Lipson D., Yakhini Z., 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48.

Efron B., Tibshirani R., 2007. On testing the significance of sets of genes. Ann. Appl. Statist., 1, 107-129.

Emilsson V., Thorleifsson G., Zhang B., Leonardson A.S., Zink F., Zhu J., Carlson S., Helgason A., Walters G.B., Gunnarsdottir S., 2008. Genetics of gene expression and its effect on disease. Nature, 452, 423-428.

Falcon S., Gentleman R., 2007. Using GOstats to test gene lists for GO term association. Bioinformatics, 23, 257-258.

Farrall M., 2004. Quantitative genetic variation: a post-modern view. Hum. Mol. Genet., 13, Spécial Issue, 1, R1-R7.

Ferrara C.T., Wang P., Neto E.C., Stevens R.D., Bain J.R., Wenner B.R., Ilkayeva O.R., Keller M.P., Blasiole D.A., Kendziorski C., 2008. Genetic networks of liver metabolism revealed by integration of metabolic and transcriptional profiling. PLoS Genet., 4, e1000034.

Georges M., 2007. Mapping, fine mapping, and molecular dissection of Quantitative Trait Loci in domestic animals. Ann. Rev. Genomics Hum. Genet., 8, 131-162.

Ghazalpour A., Doss S., Sheth S.S., Ingram-Drake L.A., Schadt E.E., Lusis A.J., Drake T.A., 2005. Genomic analysis of metabolic pathway gene expression in mice. Genome Biol., 6, R59.

Gibson G., Weir B., 2005. The quantitative genetics of transcription. Trends Genet., 21, 616-623.

Gilbert H., Le Roy P., 2003. Comparison of three multitrait methods for QTL detection. Genet. Sel. Evol., 35, 281-304.

GuhaThakurta D., Xie T., Anand M., Edwards S.W., Li G., Wang S.S., Schadt E.E., 2006. *Cis*-regulatory variations: a study of SNPs around genes showing cis-linkage in segregating mouse populations. BMC Genomics, 7, 235.

Hocquette J.F., Cassar-Malek I., Scalbert A., Guillou F., 2009. Contribution of genomics to the understanding of physiological functions. J. Physiol. Pharmacol., 60, numéro spécial, Suppl., 3, 5-16.

Hubner N., Wallace C.A., Zimdahl H., Petretto E., Schulz H., Maciver F., Mueller M., Hummel O., Monti J., Zidek V., 2005. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet., 37, 243-253.

Jacob F., Monod J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol., 3, 318-356.

Jansen R.C., Nap J.P., 2001. Genetical genomics: the added value from segregation. Trends Genet., 17, 388-391.

Jiang Z., Gentleman R., 2007. Extensions to gene set enrichment. Bioinformatics, 23, 306-313.

Kanehisa M., Goto S., Hattori M., Aoki-Kinoshita K.F., Itoh M., Kawashima S., Katayama T., Araki M., Hirakawa M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res., 34, D354-D357.

Keller A., Backes C., Lenhof H.P., 2007. Computation of significance scores of unweighted gene set enrichment analyses. BMC Bioinformatics, 8, 290.

Kim S.Y., Volsky D.J., 2005. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics, 6, 144.

Kirst, M., Myburg A.A., De Leon J.P., Kirst M.E., Scott J., Sederoff R., 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol., 135, 2368-2378.

Kruglyak L., Storey J.D., 2009. Cause and express. Nat. Biotechnol., 27, 544 545.

Lan H., Rabaglia M.E., Schueler K.L., Mata C., Yandell B.S., Attie A.D., 2004. Distinguishing covariation from causation in diabetes: a lesson from the protein disulfide isomerase mRNA abundance trait. Diabetes, 53, 240-244.

Le Mignon G., Desert C., Pitel F., Leroux S., Demeure O., Guernec G., Abasht B., Douaire M., Le Roy P., Lagarrigue S., 2009. Using transcriptome profiling to characterize QTL regions on chicken chromosome 5. BMC Genomics, 10, 575.

Lum P.Y., Chen Y., Zhu J., Lamb J., Melmed S., Wang S., Drake T.A., Lusis A.J., Schadt E.E., 2006. Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J. Neurochem., 97 Suppl 1, 50-62.

Masuzaki H., Flier J.S., 2003. Tissue-specific glucocorticoid reactivating enzyme, 11 beta-hydroxysteroid dehydrogenase type 1 (11 beta-HSD1)-a promising drug target for the treatment of metabolic syndrome. Curr. Drug Targets Immune Endocr. Metabol. Disord., 3, 255-262.

Monks S.A., Leonardson A., Zhu H., Cundiff P., Pietrusiak P., Edwards S., Phillips J.W., Sachs A., Schadt E.E., 2004. Genetic inheritance of gene expression in human cell lines. Am. J. Hum. Genet,75, 1094-1105.

Mootha V.K., Lindgren C.M., Eriksson K.F., Subramanian A., Sihag S., Lehar J., Puigserver P., Carlsson E., Ridderstrale M., Laurila E., 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately

downregulated in human diabetes. Nat. Genet., 34, 267-273.

Morley M., Molony C.M., Weber T.M., Devlin J.L., Ewens K.G., Spielman R.S., Cheung V.G., 2004. Genetic analysis of genome-wide variation in human gene expression. Nature, 430, 743-747.

Nadaf J., Pitel F., Gilbert H., Duclos M.J., Vignoles F., Beaumont C., Vignal A., Porter T.E., Cogburn L.A., Aggrey S.E., Simon J., Le Bihan-Duval E., 2009. QTL for several metabolic traits map to loci controlling growth and body composition in an F2 intercross between high- and low-growth chicken lines. Physiol. Genomics, 38, 241-249.

Pfeifer D., Kist R., Dewar K., Devon K., Lander E.S., Birren B., Korniszewski L., Back E., Scherer G., 1999. Campomelic dysplasia translocation breakpoints are scattered over 1 Mb proximal to SOX9: evidence for an extended control region. Am. J. Hum. Genet., 65, 111-124.

Ponsuksili S., Jonas E., Murani E., Phatsara C., Srikanchai T., Walz C., Schwerin M., Schellander K., Wimmers K., 2008. Trait correlated expression combined with expression QTL analysis reveals biological pathways and candidate genes affecting water holding capacity of muscle. BMC Genomics, 9, 367.

Potokina E., Druka A., Luo Z., Wise R., Waugh R., Kearsey M., 2008. Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J., 53, 90-101.

Rockman M.V., Kruglyak L., 2006. Genetics of global gene expression. Nat. Rev. Genet., 7, 862-872.

Schadt E.E., Monks S. A., Drake T.A., Lusis A.J., Che N., Colinayo V., Ruff T.G., Milligan S.B., Lamb J.R., Cavet G., 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature, 422, 297-302.

Schadt E.E., Lamb J., Yang X., Zhu J., Edwards S., Guhathakurta D., Sieberts S.K., Monks S., Reitman M., Zhang C., 2005. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet., 37, 710-717.

Storey J.D., Akey J.M., Kruglyak L., 2005. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol., 3, e267.

Stranger B.E., Forrest M.S., Clark A.G., Minichiello M.J., Deutsch S., Lyle R., Hunt S., Kahl B., Antonarakis S.E., Tavare S., 2005. Genome-wide associations of gene expression variation in humans. PLoS Genet., 1, e78.

Subramanian A., Tamayo P., Mootha V.K., Mukherjee S., Ebert B.L., Gillette M.A., Paulovich A., Pomeroy S.L., Golub T.R., Lander E.S., Mesirov J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA, 102, 15545-15550.

Subramanian A., Kuehn H., Gould J., Tamayo P., Mesirov J.P., 2007. GSEA-P: a desktop application for gene set enrichment analysis. Bioinformatics, 23, 3251-3253.

Wayne M.L., McIntyre L.M., 2002. Combining mapping and arraying: An approach to candidate gene identification. Proc. Natl. Acad. Sci. USA, 99, 14903-14906.

Williams R.B., Chan E.K., Cowley M.J., Little P.F., 2007. The influence of genetic variation on gene expression. Genome Res., 17, 1707-1716.

Yaguchi H., Togawa K., Moritani M., Itakura M., 2005. Identification of candidate genes in the type 2 diabetes modifier locus using expression QTL. Genomics, 85, 591-599.

Yamashita S., Wakazono K., Nomoto T., Tsujino Y., Kuramoto T., Ushijima T., 2005. Expression Quantitative Trait Loci analysis of 13 genes in the rat prostate. Genetics, 171, 1231-1238.

Yvert G., Brem R.B., Whittle J., Akey J.M., Foss E., Smith E.N., Mackelprang R., Kruglyak L., 2003. Transacting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat. Genet., 35, 57-64.

Zhang B., Schmoyer D., Kirov S., Snoddy J., 2004. GOTree Machine GOTM: a web-based platform for interpreting sets of interes ting genes using Gene Ontology hierarchies. BMC Bioinformatics, 5, 16.

Zhou X., Su Z., 2007. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics, 8, 246.

Zivy M., de Vienne D., 2000. Proteomics: a link between genomics, genetics and physiology. Plant Mol. Biol., 44, 575-580.

Résumé

De nombreux progrès ont été réalisés ces dernières années en génomique. Le développement de technologies à base de supports miniaturisés, permet aujourd'hui d'explorer les génomes tant au niveau de leur structure que de leur expression. Les puces à ADN permettent ainsi de génotyper plusieurs milliers de marqueurs SNP d'un génome ou encore de mesurer le niveau d'expression de plusieurs milliers de gènes d'un tissu. Combiner l'information génotypique avec des mesures phénotypiques élémentaires (ARNm, protéines ou encore métabolites) ouvre de nouvelles perspectives dans l'étude du fonctionnement du vivant et a donné naissance à un nouveau concept, la «génétique génomique». Cet article est centré sur les apports de la «génétique génomique» dans le contexte de la détection de QTL (*Quantitative Trait Locus*). Après avoir défini la notion de QTL d'expression (eQTL), cet article propose dans un premier temps un bilan des différents programmes de cartographie de QTL d'expression décrits dans la littérature. Sont ensuite présentés les différentes méthodes utilisant des données d'expression pour préciser ou caractériser fonctionnellement des régions QTL responsables de la variation de caractères d'intérêt avec des exemples concernant les animaux d'élevage.

Abstract

Contribution of Functional Genomics to the Fine Mapping of QTL

Much progress has been made in recent years in the genomics field. The development of technologies -based on miniaturized arrays makes it possible to explore genomes at both structural and functional levels. DNA microarrays allow to genotype several thousands of SNP in a genome, or measure the expression level of several thousands of genes in a tissue. Strategies combining genotypic information with elementary phenotypes (mRNA, proteins or metabolites) open new perspectives in biology research and are grouped under the new concept of «genetical genomics». We will focus here on the contributions of the «genetical genomics» in the context of QTL detection. Firstly, after defining the concept of expression QTL (eQTL), this article reports the main results on expression QTL mapping reported in the literature. Then, the different methods using expression data to refine or functionally characterize a QTL region are presented and illustrated through some examples on model and livestock species.

LE MIGNON G., BLUM Y., DEMEURE O., DIOT C., LE BIHAN-DUVAL E., LE ROY P., LAGARRIGUE S., 2010. Apports de la génomique fonctionnelle à la cartographie fine de QTL. Inra Prod. Anim., 23, 343-358.